IDENTIFYING DEMAND INDICATORS FOR AIR FORCE RECRUITING SERVICE WITH DISCRIMINANT ANALYSIS

THESIS

Jason L. Williams, 1st Lieutenant, USAF

AFIT/GOR/ENS/01M-18

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U. S. Government.

AFIT/GOR/ENS/01M-18

IDENTIFYING DEMAND INDICATORS FOR AIR FORCE RECRUITING SERVICE WITH DISCRIMINANT ANALYSIS

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Jason L. Williams, B.S.

1st Lieutenant, USAF

March 2001

AFIT/GOR/ENS/01M-18

IDENTIFYING DEMAND INDICATORS FOR AIR FORCE RECRUITING SERVICE WITH DISCRIMINANT ANALYSIS

Jason L. Williams, B.S.
1st Lieutenant, USAF

Approved:

Kenneth W. Bauer, Ph.D. (Chairman)	date
Jeffrey W. Lanning, Maj, USAF (Member)	date

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Dr. Bauer, for his flexibility and guidance throughout the course of this thesis effort. The insight and experience was certainly appreciated. Major Jeff Lanning also provided invaluable constructive criticism and helped me step away from the problem and see the big picture. Also I would like to acknowledge the help I received from Captain Frank O'Fearna from Air Force Recruiting Service for both the support and latitude provided to me in this endeavor.

I am also indebted to my friends and family for their support during this effort. My peers at AFIT were an excellent source for advise and were there when I just needed someone to listen. Mostly I would like to thank my wife who was my best source of support even though she was also completing a challenging academic program simultaneously.

Jason L. Williams

Table of Contents

List of Figures

Figure Page

List of Tables

Table Page

Abstract

A changing public disposition towards military service has all four military branches rethinking recruiting practices. This Air Force is reacting to this new recruiting climate by increasing the bonuses for new recruits, pumping up advertising budgets, and bolstering recruiting personnel levels.

This thesis provides a new tool for assessing how to allocate these new resources. Discriminant Analysis is used to identify population characteristics that categorize recruiting locations. A methodology is constructed that can discriminate between communities where interest is high in military service and where recruiting efforts will not be productive.

Identifying Demand Indicators for Air Force Recruiting Service With Discriminant Analysis

I. Introduction

1.1 Statement of Problem

The difficulty of finding new recruits to join the military has been well documented. Hafemeister reports a 1999 shortfall in recruiting alerted Air Force leaders that recruiting needs to be a top priority. The Air Force has three initiatives that will combat the problem: more recruiters, more money for advertising, and more bonuses (2000). These are positive steps that are targeted to meet the needs of the Air Force's *customers*, people willing and eligible to join the Air Force. Focusing on the customer is the hallmark of process improvement (Snee, 1996). Now that Air Force Recruiting Service has more resources to meet its customer's needs it has the responsibility to use those resources effectively.

In 1998 the Air Force employed 890 recruiters. The 1999 shortfall prompted Air Force leaders to increase the number of recruiters. By the end of the year 2000 there were 1,350 recruiters in the field and the Air Force recruiting goal was met (Rolfsen, 2000). Three hundred more recruiters are coming in the year 2001 to meet the increasing recruiting goals of the future. Air Force Recruiting Service has not had to deal with this level of prolific manning changes in several years. When the new recruiters were assigned to Air Force Recruiting Service there was no policy in place to guide their placement.

This research employs a multivariate statistical technique, discriminant analysis, to the field of military recruitment for the purpose of increasing decision makers understanding of the customer base. The purpose of this thesis is to find a practical method for linking demographic indicators to recruiting databases and employing discriminant analysis to discover customer attributes.

1.2 Literature Review

The literature review covers four areas: previous research on community influence, an overview of Air Force Recruiting Service structure and practices, a explanation of the data that was used, and background on performing a discriminant analysis study.

The purpose of this thesis is to identify indicators of the demand in the Air Force. We searched for those indicators by looking at the communities recruits came from rather than economic statistics from individuals. It is the premise of this thesis that people are greatly influenced by the communities they live in and that there exist community attributes that indicate a higher propensity to join the Air Force. The first part of the literature review discusses previous research on how the community influences the individual. The second part of the literature review explains how Recruiting Service structure and practices affect the thesis research. Air Force Recruiting Service has built procedures and practice over several decades. Any research effort about Air Force recruiting requires an understanding of these procedures and practices. The third section of the literature review explains the data that was made available by Recruiting Service and the data obtained from outside sources. Recruiting Service made available a database that contains information about recruit attributes. These recruit attributes are very useful for defining groups for study. The Air Force does not collect or maintain demographic

information about the places recruits come from. Demographic data about communities was obtained from a general commercial source named eNeighborhoods. This is a computer program widely used in the realty industry to provide customers with accurate and comprehensive information about the neighborhoods they want to live in.

The last section of the literature review is an overview of current discriminant analysis theory. There are three areas of theory covered: underlying assumptions, classification functions, and assessing results. The underlying assumptions of discriminant analysis are that the groups are multivariate normal and have equal covariance matrices. This part of the literature review explains how to evaluate data adherence to these assumptions and steps to take if the assumption are not met. There are several different methods of classifying data when performing discriminant analysis. Two methods, Fisher's method and the quadratic discriminant method, are employed in this thesis. The literature review explains how the two discriminant functions classify the data and how a researcher can choose the appropriate function. Once a classification function has been generated the researcher needs tools that assess classification accuracy and the importance of the independent variables that make up the function. The literature review introduces tools commonly used for this purpose and how to interpret their results.

The literature review introduces previous research that motivates the direction of this research, explains the data that was available to accomplish the research, and summarizes current tools used in the field of discriminant analysis. The methodology chapter describes how that theory was applied to the specific circumstances of this thesis.

1.3 Methodology

The methodology chapter covers how groups were formed with Recruiting Service data, how the Recruiting Service data was linked to demographic data, and how discriminant analysis theory was applied in this thesis.

Significant effort was spent defining groups for this study. Group definition is important because it frames the research question. If this is done incorrectly, the research results can be applied incorrectly. In the methodology section, different approaches to defining groups are discussed.

Collecting demographic data can be expensive and time consuming. A methodology that minimized the difficulty of collecting data without compromising accuracy was established. The final section of the methodology section covers specific discriminant analysis details. Specific methodologies for determining variable importance, variable reduction, and result verification are explained in this section. The fourth chapter presents the results of applying this methodology.

1.4 Results

The results research resulted in identifying indicators of demand conducive to enlisting in the Air Force. These indicators are economic and demographic variables that allow recruiters to objectively compare potential recruiting areas. More importantly this thesis built and tested a methodology that can be used to answer future Air Force Recruiting Service problems.

This chapter shows the groups that were formed and explains which classification method was used and why. The resulting classification function is explained and the impacts of the variables that are used to formulate it are described.

The business of military recruiting has always been difficult. This research explored and tested a methodology that will potentially improve planning efficiency at several levels of the recruiting process.

II. Literature Review

The literature review gives detailed information on subjects pertinent to this research effort. The first section discusses research concerning how the community a potential recruit lives in can influence the individual and how this thesis will use that to benefit the Air Force. The second section gives a review of Air Force recruiting structure and procedures. The third section introduces the data that has been made available for the study and explains how it was used. The fourth section is a detailed explanation of the statistical formulas that were used to perform discriminant analysis.

2.1 Community Influencing the Individual

The community that a person lives in influences several aspects of their lives. Research efforts have uncovered community factors that influence individual behavior. In a study about child development, Laura DeHaan (1999) found that "an individual family's income situation wasn't associated with (adolescent) risk taking, but the community's economic status was". DeHann's research extended beyond economic indicators, another interesting finding was that the community's attitude about adult drinking influenced adolescent drinking the delinquent behavior of youth. In a separate study Vanfossen (1996) linked occupational distributions, housing types and transportation networks to community violence. These studies emphasized collecting data and did not use sophisticated techniques for evaluation.

There is also literature that suggests that the community attitudes affect the quality of school systems. A major factor a parent considers when purchasing a house is the characteristics of the school that their children will attend. The article "The Eight Key Questions Sellers Hope Buyers Never Ask" says homebuyers search for the highest-

quality school district they can find (2000). Lu Battaglieri, the president of the Michigan Education Association, adds, "The success of the community is directly tied to the quality of its public schools" (O'Conner, 2000). Given this premise, characteristics of high schools can be found that identify demographic groups that are more likely to enlist in the Air Force.

This premise of this thesis is that some communities influence their members to join the military at a greater degree than other communities. This research will concentrate on identifying economic and social indicators that can be used to distinguish these communities.

2.2 Recruiting Structure

Air Force Recruiting Service is located at Randolph Air Force Base in San Antonio Texas. It divides its responsibilities into a hierarchical relationship that can be seen in Figure 1. Air Force Recruiting Service has relied on about 900 field recruiters to accomplish its mission (O'Fearna, 2000). The Air Force increased the number of recruiters to 1350 at the end of 2000 with plans in place that will increase the number to 1650 within the year. Recruiter placement is handled in bottom-up fashion. The flight commanders have the responsibility of selecting a field office for a gain or loss of a billet. Once a request is made it is coordinated through each level until all requests for a given year are approved at the Air Force Recruiting Service level.

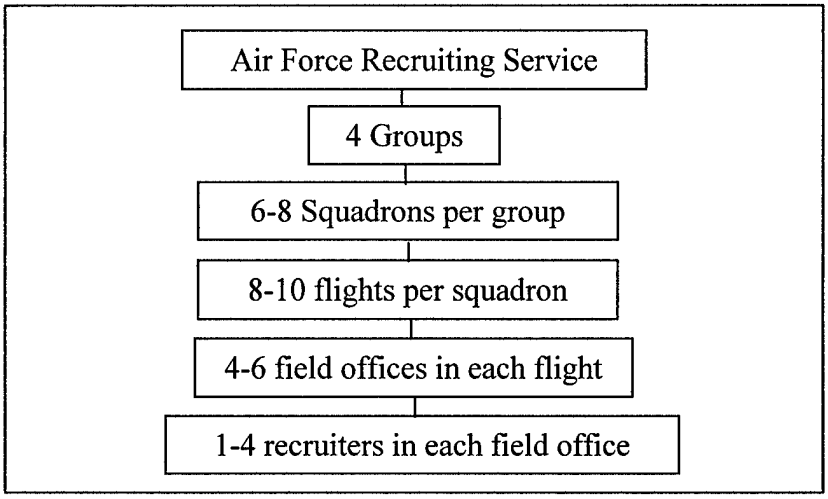


Figure 1 Air Force Recruiting Service Structure

Field office locations and manning levels have been rather static. When the Air Force added new recruiters there was not a procedure in place to identify where the recruiters would be most useful (O'Fearna, 2000). This thesis identifies measures of merit that can identify cities that are conducive for recruiting.

There are a lot of factors that go into the recruiting process. A recruiter's duties can be separated into four different categories: planning, prospecting, selling, and processing as seen in Figure 2 (Cordeiro and Friend, 1998). Planning is devising strategies that will meet short and long-term recruiting goals. This game plan is made up of step-by-step processes that accomplish the other three duty categories. Successful recruiters will have a plan in place before prospecting or selling and will make constant revisions as they gain experience.

Prospecting is informing people about the Air Force. There are four methods of prospecting: face-to-face, telephone, referrals, or walk-ins. All four methods must be used together to build interest in a prospect. Sometimes a motivated prospect walks-in

and wants to join right away. Most of the time the recruiter has to work very hard to build enough interest to even get the prospects to come to the office.

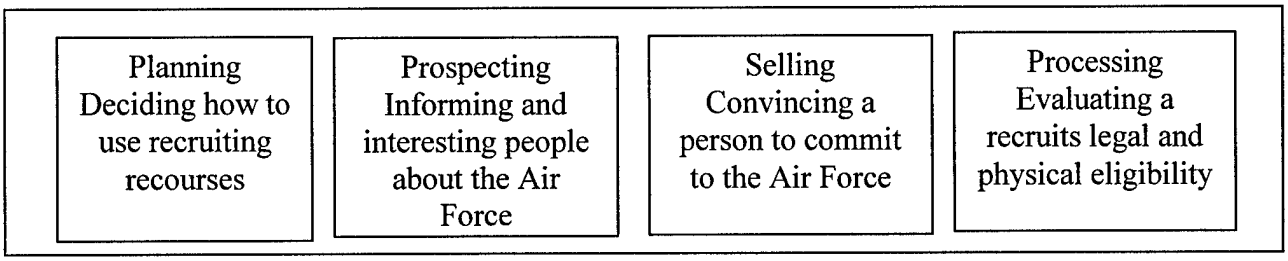


Figure 2 The Four Steps of Recruiting

After a recruiter builds initial interests in a prospective recruit, what the Air Force has to offer must be sold to the recruit. The sales phase is felt to be the key element in the recruiting phase because an applicant is "won or lost" in this phase. Processing is the formal evaluation of a recruit. These evaluations include criminal background checks, aptitude testing, and physical testing. During this phase the recruit has the freedom to change their mind and decide not to enlist. This means the recruiter continues the sales process until the prospect has arrived at boot camp. Sales and processing are time consuming processes. Therefore if planning and prospecting took less time recruiters could spend more time selling (Longhorn, 2000).

2.3 Available Data

Recruiting Service maintains information about the people that have enlisted in the Air Force. Most of the information is data required for the enlistment into the Air Force. Every recruit is required to meet minimum standards on the Armed Services Vocational Aptitude Battery or ASVAB test. The ASVAB has a 100-point scale and recruits must score a minimum of 40 points to be eligible to join the Air Force. The three

other services have lower requirements and sometimes establish requirement waivers in order to meet recruiting shortfalls. General Michael Ryan, the Air Force chief of staff, has made it clear that the Air Force will not lower standards even if it means recruiting goals are not met (Palmer, 2000).

Gender and race information for each recruit is also included. It is very important to the Air Force that it does not discriminate against any race or gender. By maintaining and periodically monitoring this data, the Air Force can keep track of the cultural demographics of its workforce.

The Air Force does not maintain economic or demographic data on the places that supply recruits. The realty industry is able to provide this sort of information for its customers on a routine basis. A leading product that realtors use is eNeighborhoods made by Iplace Professional Services. Century 21, Remax, and Coldwell Banker realtors use eNeighborhoods to provide up-to-date information on city and neighborhood demographics, public and private schools, and many other areas (eNeighborhoods, 2000).

2.4 Discriminant Analysis Overview

Performing discriminant analysis can be summed up in a few simple steps:

1. Check for multivariate normality. This is a requirement for almost all of the underlying theory of discriminant analysis.
2. Test to see if covariance matrices (Σ_i for i = 1,2) are equal. A common method, Fisher's two-group discriminant analysis, makes this assumption. If this requirement is not met, there are other methods that can be used.
3. Choose a method and compute the discriminant function to generate discriminant scores.

4. Validate the chosen method.

2.4.1 Assessing Multivariate Normality

The first step in assessing multivariate normality is to assess the univariate, or marginal, normality of each independent variable (Andrews, 1974). This can be accomplished with several tools including likelihood tests and normal probability plots. A complete discussion of these techniques can be found in Neter (1996). Although marginal normality does not imply multivariate normality, the presence of most types of deviation from normality will be revealed in the univariate analysis (Andrews, 1974).

If problems are found in the marginal normality, the best solution is to apply a transformation that addresses the specific problem. The most common transformations are exponential and logarithmic transformations for increasing or decreasing variance (Neter *et al*, 1996).

2.4.2 Testing for Unequal Covariance Matrices

The covariance matrix for a multivariate data set is analogous to the variance statistic for a single variable (Giri, 1996). The covariance matrix, labeled Σ, is rarely known for real world data. The sample variance, shown in Equation 1, is used to approximate Σ.

$$\mathbf{S} = \frac{1}{n-1} \cdot \mathbf{X}_d^T \cdot \mathbf{X}_d \text{, where } \mathbf{X}_d = \mathbf{X}_i - \begin{bmatrix} 1 \\ \cdots \\ 1 \end{bmatrix} \cdot \overline{\mathbf{X}_i}^T \tag{1}$$

The Bartlett and Box test for unequal covariance matrices is used in this research to determine whether a pooled covariance matrix can be used or not. Giri (1996) and

Bauer (2000) show formulations of this test for testing the equality of multiple covariance matrices. The following is a simplified formulation for use in a two-group problem applicable to this research effort.

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be matrices of independent variables corresponding to two *a priori* defined groups. Now define $\mathbf{A}_i$ to be the crossproduct of the mean corrected data for i = 1, 2.

$$\mathbf{A}_i = \left[\mathbf{X}_i - \begin{bmatrix}1\\ \cdots \\ 1\end{bmatrix} \cdot \overline{\mathbf{X}_i}^T\right] \cdot \left[\mathbf{X}_i - \begin{bmatrix}1\\ \cdots \\ 1\end{bmatrix} \cdot \overline{\mathbf{X}_i}^T\right]^T \quad i = 1,2 \qquad (2)$$

Also Define

$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$
p = # of variables
n_i = sample size in group I
$N = n_1 + n_2$
$n = N - 2$

The test for equal covariance matrices is a standard hypothesis test. A test statistic is found and is compared to a χ^2 distribution where α is the chance of type I error and with $1/2 \cdot p \cdot (p-1)$ degrees of freedom.

Null Hypothesis: $\Sigma_1 = \Sigma_2$

Rejection Region: test statistic > $\chi^2_{1-\alpha,\frac{1}{2}\cdot p\cdot(p-1)}$

The test statistic is

$$\omega 2 = -2 \cdot \rho \cdot \ln(W) \qquad (3)$$

Where

$$\rho = 1 - \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n} \right) \cdot \frac{(2 \cdot p^2 + 3 \cdot p - 1)}{6 \cdot (p+1) \cdot (q-1)}$$

$$W = e^V \cdot \left[\left(\frac{n}{n_1 - 1} \right)^{\frac{(p \cdot (n_1 - 1))}{2}} \cdot \left(\frac{n}{n_2 - 1} \right)^{\frac{(p \cdot (n_2 - 1)}{2}} \right]$$

$$V = \left[\frac{(n_1 - 1)}{2} \cdot \ln(|\mathbf{A}_1|) + \frac{(n_2 - 1)}{2} \cdot \ln(|\mathbf{A}_2|) - \frac{n}{2} \cdot \ln(|\mathbf{A}|) \right]$$

These are complicated formulas that can be difficult to compute. Guri (1996) shows a much simpler approximation for the test statistic derived by Box in Equation 4. ρ is the term defined in Equation 3 and S_i is the covariance matrix of the ith group and S_p is the pooled covariance matrix defined in Equation 5:

$$\rho \cdot \left[n \cdot \ln(|\mathbf{S}_p|) - (n_1 - 1) \cdot \ln(|\mathbf{S}_1|) - (n_2 - 1) \cdot \ln(|\mathbf{S}_2|) \right] \tag{4}$$

$$\mathbf{S}_p = \frac{1}{n_1 + n_2 - 2} \cdot \left[(n_1 - 1) \cdot \mathbf{S}_1 + (n_2 - 1) \cdot \mathbf{S}_2 \right] \tag{5}$$

2.4.3 Classification Methods

One goal of a classification method is to assign a scalar score to each object in the data set. The score determines class membership. Two classification methods are covered in this section, Fisher's linear approach and the quadratic discriminant function. Fisher's approach has a body of supporting literature and is simple to compute. The quadratic discriminant function is a more robust classifier in that it does not require equal covariance matrices.

Fisher's method finds a linear combination of the object's attributes with the goal of maximizing the distance between the means and minimizing the variance (Dillon, 1984). Fisher proved that Equation 6 forms such a linear combination.

$$\hat{b} = \mathbf{S}_p^{-1} \cdot (\overline{\mathbf{X}_1} - \overline{\mathbf{X}_2}) \tag{6}$$

Where S_p is the sample pooled covariance matrix.

This equation produces a vector of weights. Scores for each object are calculated by multiplying each object attribute by the appropriate weight. An entire group's scores can be calculated by calculating Equation 7.

$$score_i = \mathbf{X}_i \cdot \left(\hat{b} \right), \text{ for group I} \tag{7}$$

After scores have been calculated, a classification rule is imposed. For Fisher's method, the simplest rule is using the midpoint of the scores as a dividing point. First determine the mean score for both groups and then determine the overall mean of the scores. Call the group with mean score smaller than the overall mean group A and the group with score larger than the overall mean group B. New objects with scores smaller than the overall mean are classified as group A and objects with scores larger than the overall mean are classified as group B. This classification method will not necessarily achieve 100% accuracy. The misclassification error percentage as calculated by applying the discriminant function to the data that was used to formulate the weights is called the apparent error-rate. The following is a step-by-step guide to determining the apparent error rate for Fisher's method. This method only works for groups with equal number of objects.

1. Determine the means of the scores for groups 1 and 2.
2. Find the mean of the combined scores using Equation 8.

$$Midpoint = .5 \cdot (\overline{\mathbf{X}_1} - \overline{\mathbf{X}_2})^T \cdot \mathbf{S}_p \cdot (\overline{\mathbf{X}_1} + \overline{\mathbf{X}_2}) \tag{8}$$

3. Determine scores for $\mathbf{X}_1$ and $\mathbf{X}_2$ using Equation 7.
4. Let the number of correctly classified objects for $\mathbf{X}_1$ be c_1 and likewise let the number of correctly classified objects for $\mathbf{X}_2$ be c_2.
5. Calculate the apparent error-rate (Aper) the ratio of misclassified objects to the total number of objects and can be seen in Equation 9.

$$Aper = \frac{(n_1 - c_1) + (n_2 - c_2)}{n_1 + n_2} \tag{9}$$

One advantage of using Fisher's method is the existence of a statistical test to assess the Mahalanobis distance between the means of the groups when the covariance matrices of the two groups are equal (Giri, 1996). This allows for a quick test of data for the likelihood of successful discriminant analysis. The following is a description of Hotelling's T^2 statistical test:

Null Hypothesis: $\mu_1 = \mu_2$

Rejection Region: Test statistic > $F_{(1-\alpha, n_1+n_2-p-1)}$

Test Statistic:
$$\frac{n_1+n_2-p-1}{p\cdot(n_1+n_2-2)}\cdot T^2 \tag{10}$$

Where $$T^2=\frac{n_1\cdot n_2}{n_1+n_2}\cdot\left(\overline{\mathbf{X}_1}-\overline{\mathbf{X}_2}\right)^T\cdot\mathbf{S}_p{}^{-1}\cdot\left(\overline{\mathbf{X}_1}-\overline{\mathbf{X}_2}\right)$$

Quadratic Discriminant Scores retain the underlying assumption of multivariate normality but allows for groups with unequal covariance matrices (Dillon, 1984). Another advantage of quadratic discriminant scores (D_q scores) is that it can separate groups that are not linearly separable (Bauer, 2000). The disadvantage of using D_q scores is the higher level of computational complexity.

D_q scores are an approximation of the natural log of the likelihood estimator. Each object classified receives a score from the likelihood estimator for the first group and a score from the likelihood estimator for the second group. The formula for calculating the quadratic discriminant scores for an object i from matrix X is,

$$\begin{aligned} D_1^q(X_i)&=-\frac{1}{2}\cdot\ln|\mathbf{S}_1|-\frac{1}{2}\cdot(X_i-\overline{X_1})^T\cdot\mathbf{S}_1^{-1}\cdot(X_i-\overline{X_1})+\ln(\frac{n_1}{n_1+n_2})\\ D_2^q(X_i)&=-\frac{1}{2}\cdot\ln|\mathbf{S}_2|-\frac{1}{2}\cdot(X_i-\overline{X_2})^T*\mathbf{S}_2^{-1}\cdot(X_i-\overline{X_2})+\ln(\frac{n_2}{n_1+n_2}) \end{aligned} \tag{11}$$

After both scores have been calculated the object is classified into the group with the larger D_q score (higher likelihood).

Quadratic discriminant scores are similar to Fisher discriminant scores in that they will not necessarily correctly classify 100% of all objects. An apparent error rate can be calculated with Equation 9.

2.4.4 Deciding on an Classification Method

Deciding on a classification method is a straightforward decision. If we can take the covariance matrices as equal, Fisher's method is appropriate, but if not, we will use quadratic discriminant scores. Unfortunately, application may not prove that straightforward. Some groups with unequal covariance matrices can be correctly classified by Fisher's method if the means of the groups are far enough apart. This is seen in figure 3.

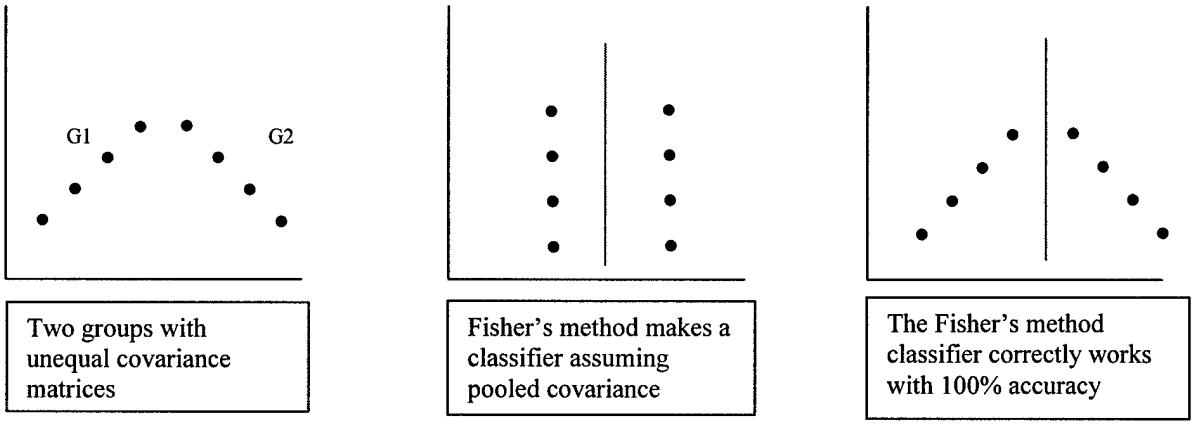


Figure 3 The Flexibility of Fisher's Method

We recommend using both methods to classify the data instead of depending solely on the condition of the covariance matrices to guide the decision. Incorporate into the decision each method's performance by measuring the apparent error rate and the error rate on holdout data.

2.4.5 Variable Contribution

When a discriminant function derives a classification function, it makes use of all the data available. Discovering which variables make the greatest contribution to classifying the data is the main focus of this research. The score generated by the discriminant function is an artificial variable that, by design, is the most effective tool for

distinguishing between the defined groups. The independent variables that have the highest magnitude of correlation with the discriminant scores, arguably, have the most contribution to classifying the data. Discriminant loadings are the correlations of the independent variables with the scores produced by the chosen discriminant function (Dillon, 1996). Figure 4 is an illustration that shows how variables can exhibit varing influence on classification.

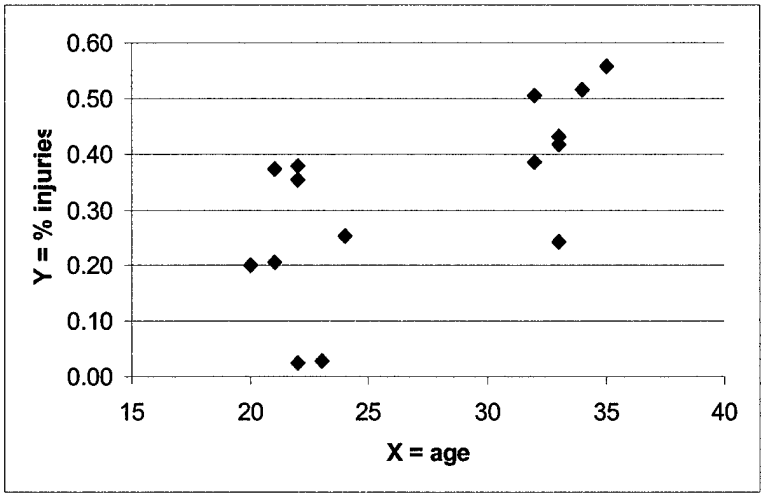


Figure 4 Understanding The Need For Loadings

In this example, two variables have been collected by the researcher, customer age, X, and percentage of injuries, Y. Fisher's method produces the classification function $-7.7 \cdot X - 4.0 \cdot Y$. Because the magnitude of the coefficents are simmilar, a researcher may infer that both variables are important for classification. This is a false assumption that is caused by the different scales of the two variables. Discriminant loadings give a more accurate assesment because they are a correlation measurement and therefore unitless. The loadings for this problem show a correlation of -.999 for X and -.189 for Y

with the discriminant scores. Age is clearly the more important variable for classifing and the loadings reflect that.

When computing discriminant loadings for groups with equal covariance matrcies, as is the case in Fisher's method, Equation 2.12 will produce a vector of the correlations of each variable with the discriminant function.

Define

$$\mathbf{D}_x = \begin{bmatrix} \frac{1}{\sqrt{S_{i,i}}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sqrt{S_{i,i}}} \end{bmatrix}$$

$$b = S^{-1} \cdot (\overline{X_1} - \overline{X_2})$$

$$D_{bx} = \left(b^T \cdot S \cdot b\right)^{\frac{1}{2}}$$

$$loadings = D_{bx} \cdot D_x \cdot S \cdot b \qquad (12)$$

If the groups do not have equal covariance matrices loadings can still be computed. Computing a univariate correlation of each variable with the discriminant scores one at a time with Equation 13 provides a loading vector.

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \cdot \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}} \qquad (13)$$

This chapter presented past research that supports the position that the community influences the individuals that live within it and therefore traits about communities can be found that predicts the actions of individuals. This chapter also explained the data that was made available by Air Force Recruiting Service and the data that was obtained from commercial sources. The final section of this chapter gave a detailed description of the current theories of discriminant analysis. Chapter III will explain how that theory was applied to the specific criteria of this thesis.

III. Methodology

This chapter explains the data available from Air Force Recruiting Service and the method of collecting demographic indicators. In the first section, the process of defining suitable groups is explained. The second section discusses how two separate databases were used in this study. The Air Force Recruiting Service database was used to define the independent groups. The other database is a commercial realty product called eNeighborhoods. eNeighborhoods is used to provide demographic data about state, city, and zip codes.

3.1 Defining A Priori Groups

An *a priori* group is a collection of two sets that are mutually exclusive and exhaustive. The term "Mutually exclusive and exhaustive" implies that each object that is studied is assigned to one and only one of the groups (Dillon, 1984). Figure 5 is an example of the easiest way to do this: define groups with an indicator variable.

A car dealer wants to know the difference between male buyers and female buyers for the last year. Each object being studied, the buyers, can be assigned to one and only one group.

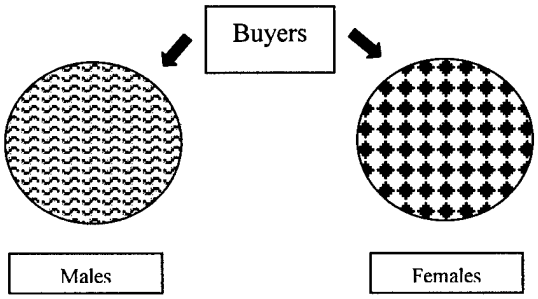


Figure 5 Groups Defined by Indicator Variable

A priori groups can also be defined on continuous variables. The main difference is that using continuous variables requires the person conducting the study to have a more active role in defining the groups. There are two ways this can be done, choosing a dividing point, or studying ranges of data. A dividing point is a single number within the continuous range of the variable that the user feels is relevant. Figure 6 is an example of defining groups on an entire continuous set.

A car dealer wants to know the difference between the buyers that spend less than the median price and those that spend more than the median price.

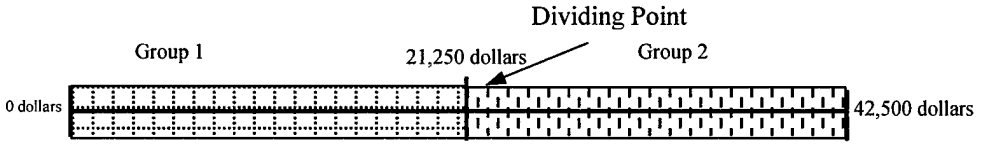


Figure 6 Defining Groups on an Entire Continuous Set

Studying ranges of the data is similar to choosing a dividing point with one major difference. This method excludes objects that could be assigned a value in the continuous spectrum of the variable in order to gain understanding about subsets of the continuous variable. In Figure 7 the car dealer from the previous example modifies the study by using subsets of a continuous variable.

A car dealer sells three price categories and wants to advertise the economy cars and the luxury cars. The economy cars range from 10,000 to 16,000 dollars and the luxury cars range from 32,000 to 42,500 dollars.

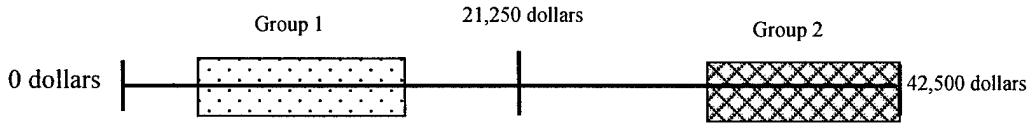


Figure 7 Defining Groups on Subsets of a Continuous Variable

A wide variety of problems can be addressed with discriminant analysis. Knowing which grouping method to use requires a careful study of the data and a firm grasp of the problem that needs to be solved. Group definition is a flexible process as long as the groups that are formed are mutually exclusive and exhaustive.

3.2 Defining Groups to Maximize Group Distinction

Here a specific methodology is employed to define groups from a continuous variable such that the two groups will be truly distinct from each other. When two groups are defined from a continuous variable using a single division point there is a set of the data that can be ambiguous as seen in the example in Figure 8. Sometimes classifying these ambiguous points is the goal of the research. In that case incorporating this data set is essential.

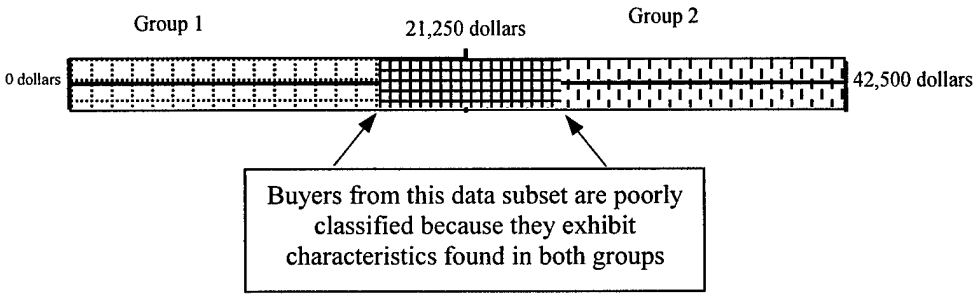


Figure 8 Ambiguous Data Points

Our research goal is identify the best and the worst places to recruit. Therefore, only data that is from high demand areas and low demand areas is included in the research. This thesis uses a six-step process to limit the effects of ambiguous data on the classification process, which is presented in Figure 9.

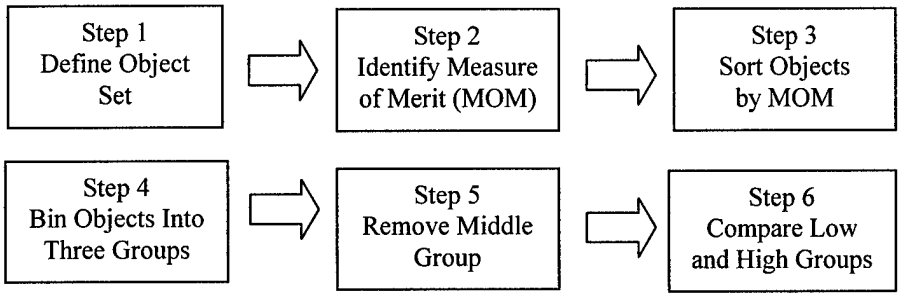


Figure 9 Removing Ambiguous Data

The first step of the process is to decide which objects will be used to generate the discriminant function. The level at which the objects are restricted depends on what question the research is meant to answer. The resulting discriminant function will be a better classifier for the data that is included and less effective for the data that has been excluded. Figure 10 is an example of the different levels at which data could be restricted for a study about breeding dogs. If the researcher wants to apply the results to the entire species the data should include all dogs. If the results are not going to be applied to dogs that weigh less than 40 pounds then data from that group does not need to be included.

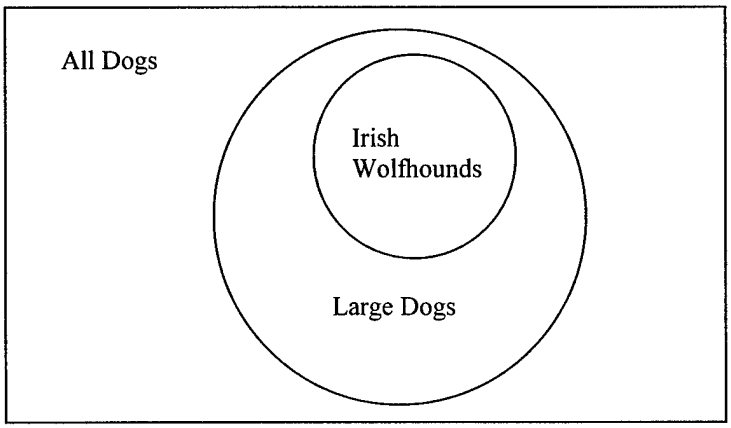


Figure 10 Restricting Study Data

The second step is to define the measure of merit (MOM) that will define the groups. The measure of merit is a continuous variable that can be associated with each object and be useful for grouping objects. The measure of merit should define groups that have a comparable number of objects. The third step of the process is to sort the objects by the MOM from least to greatest. This places all of the objects on a continuous scale so the researcher can evaluate were appropriate dividing points can be set to define the groups.

The fourth step of the process is to define two dividing points that will bin the data into three groups. These dividing points should be set with the objectives of limiting the affect of ambiguous data and producing low and high groups that are of similar sizes. The fifth step of the process is to remove the middle group of the data from the defining data set, which is illustrated in Figure 11. The sixth, and final step is to generate a discriminant function with data in the low and high groups.

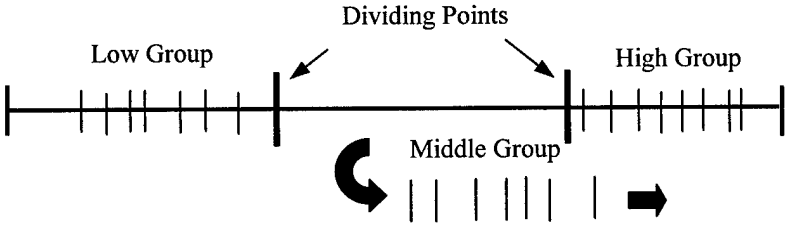


Figure 11 Removing Ambiguous Data

3.3 Air Force Recruiting Service Database

The Air Force Recruiting Service provided a database that contains data about people who have chosen to enlist in the Air Force (referred to as recruits) over the last six years. The Air Force maintains data on recruits that are gathered during their enlistment into the military. One variable, the zip code of the recruit's residence at the time of enlistment, is of particular interest to this study. Over the last six years over 185,000

people have enlisted into the Air Force from all over the country. The zip codes in this database reveal at a high-resolution level exactly where those recruits come from.

This last six years of data was made available for this research. There are two reasons to use this many years of data: lowering the variability due to individual recruiters and limiting the influence of outliers in the data. Recruiters have different personalities and different skill levels. Some of the variation between the numbers of recruits that come from different areas is due to the skill of the recruiters involved. Using six years of data insures that multiple recruiters have had responsibility for the areas that are being studied. The second reason for using six years of data is to limit the effect of a point estimate on the results. If only one year of data is used there is a high probability that some areas will not show recruit production. This would cause some areas incorporated in the study to be misclassified. Using six years of data is like taking six samples. This will smooth out the data and reduce the possibility of an outlier affecting the data.

The Recruiting Service database contains data from all over the country and some overseas locations. In order to gain meaningful insight subsets of the data must be used that are appropriate to current problem. This thesis limits the data to the 100 largest cities in the United States for three reasons. First, this limits the scope of the research to a manageable workload. Finding relevant demographic data is a time consuming process (Klopper, 2000) and if the study is too large, a timely analysis is not possible. Second, the selected cities form a good mixture of different geographical locations. Third, choosing cities with similar population sizes makes comparisons across cities more meaningful. Comparing farming villages to major cities would add confusion rather than

insight to this research effort. A list of the cities included in this study can be found at Appendix 1. This thesis identifies demand at two levels, across entire cities and within cities. Understanding how to identify a city that is favorable for recruiting aids the Air Force at the strategic level. This is information that will aid headquarter level planners in deciding on how to allocate recruiting resources. The ability to identify demand within a city will aid field recruiters at the tactical level. This will allow a recruiter to decrease planning and prospecting time by identifying potential hotspots within the recruiter's area of responsibility. These situations were investigated in the same fashion.

The first step of identifying demand across cities is to decide on an object set. The 100 largest cities in the United State were all included as objects to identify demand across cities. The second step is to identify a measure of merit that can be assigned to each city and identify the demand to join the Air Force that exists in each city. A count of the number of recruits that came from each zip code within each city was used. Discovering this number required reducing the data from the Recruiting Service database. Over 185,000 different recruits enlisted in the Air Force in the previous six years. The 34,102 recruits that came from the 100 largest cities were included in the study. The last step was to count the number of recruits that came from each city. This count data is the measure of merit that was used to identify cities that demonstrate a high demand to join the Air Force. This method of collapsing the data is illustrated in Figure 12.

The third step of defining groups was to sort the objects by the measure of merit. A table that shows the number of recruits that came from each city can be found in appendix 1. The fourth step of the process was to define two dividing points that split the

city into three groups. The cities were assigned to one of three categories: low producers, middle producers, and high producers.

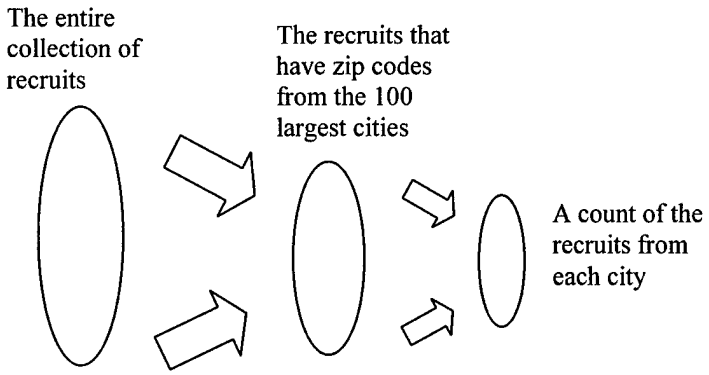


Figure 12 Defining a Measure of Merit from the Recruiting Service Data

Deciding on the dividing points that defined the group was a decision based on two objectives. The objective was to find dividing points that were natural breaks in the data that also produced groups with comparable sizes. Table 1 shows how the data was grouped. All cities with less than 200 recruits over the last six years are in the low producers group. Cities that had between 200 and 400 recruits are in the middle producers group. Finally, the cities that had over 400 residents join the Air Force over the last six years are in the high producers group.

Table 1 Grouping Cities by Number of Recruits

City	Alexandria, VA	Yonkers, NY	Glendale, CA	...	New York, NY	Akron, OH	Fort Wayne, IN
Recruits	32	35	45	...	189	195	196

Low Producers – 30 cities

City	Aurora, CO	Baton Rouge, LA	Denver, CO	...	Wichita, KN	Minneapolis, MN	Buffalo, NY
Recruits	207	212	213	...	378	378	394

Middle Producers – 40 cities

City	Austin, TX	Dallas, TX	Indianapolis, IN	…	Chicago, IL	Houston, TX	San Antonio TX
Recruits	400	402	410	…	964	1102	2161

High Producers – 30 cities

The high producers group is made up of cities where demand to join the Air Force is high and the low producers group are the cities that have low demand. The middle producers are cities that have an ambiguous demand level. These cities were not included in the analysis for the purpose of maximizing the differences between the groups. Now, as displayed in Figure 13, there are two distinct groups with an equal number of objects and the data can be dealt with as a two-group discriminant analysis problem.

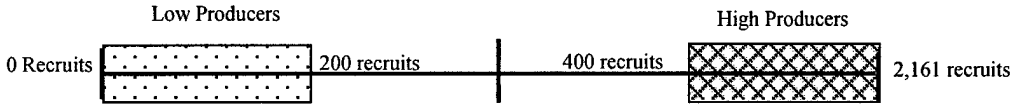


Figure 13 Two Group Problem

The methodology for defining groups for identifying demand within cities also follows the six-step process laid out in Section 3.2. The first step is to determine the object set that will define the discriminant function. This part of the study identifies locations within cities that are better for recruiting. Neighborhoods found within the 15 largest cities in the United States were used in this study. Within these large cities a zip code represents a small geographical region. Each zip code represents a neighborhood of the city. Eight neighborhoods from each city were included in this research. An example of this can be seen in Figure 14 for the city of Houston. The second step of the process is to define a measure of merit. The count of recruits that came from each zip code is the

measure of merit for this study. It is very similar to the MOM used for identifying demand across cities. In that study the number of recruits for each zip was totaled in order to find the number of recruits for the whole city. In this study there is no need to total the zip code counts because the focus is on individual neighborhoods.

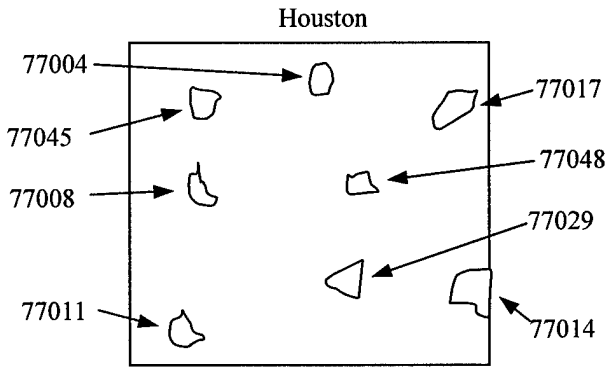
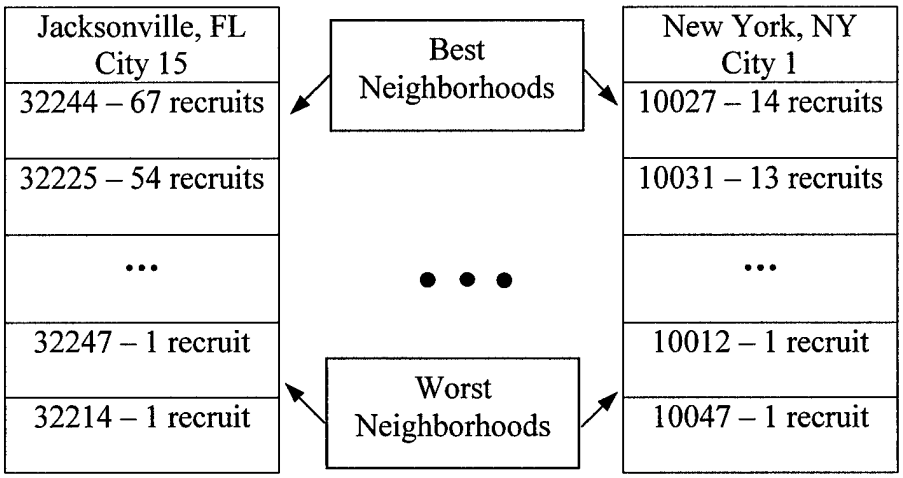


Figure 14 Zip Code Neighborhoods

The third step of the process is to sort the study objects by the measure of merit. This was done differently for this part of the study. Rather than combine all of the data from each city, the individual neighborhoods in each city were ordered, shown in Table 2.

Each of the fifteen cities had a different level of demand. This puts each city on a different scale. If all the neighborhoods were placed in the same set and sorted the best overall neighborhoods would be compared to the worst overall neighborhoods. The goal of this thesis is to compare the best neighborhoods of each city to the worst neighborhoods of each city.

Table 2 Sorting Neighborhoods Within Cities



The fourth step involves removing ambiguous data by assigning two dividing points that split the data into three groups. Each city was assigned different dividing points. The dividing points were chosen so that no more than twelve percent of the zip codes for the entire city were included in the bottom group and no more than twelve percent of the zip codes were included in the top group. The groups defined for each city had different sizes. To insure equal representation from each city four neighborhoods were randomly selected from each group for the study. An is illustration of the results of this random selection is shown in Figure 15.

The fifth step of the data reduction process is to remove the middle group from the data set. In addition to removing the middle group from the data set the neighborhoods that were not selected from the bottom and top groups were also removed from the data set. The sixth step is to compare the low and high groups. The high group was made up of all the high interest neighborhoods from all fifteen cities and the bottom group is made up of all the low interest neighborhoods from all fifteen cities.

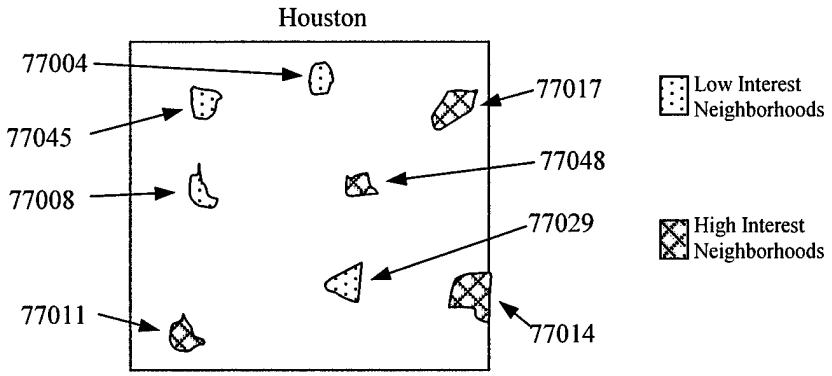


Figure 15 Neighborhoods Used for Study

3.4 Collecting Independent Variables with eNeighborhoods

eNeighborhoods is a commercial software product used by realtors to provide demographic data to home buyers. Realty firms use it because it provides accurate census and local data for cities, census tracts, and individual homes. The software is designed to provide an individual homebuyer with a report that makes comparing potential purchases an objective process. The product's depth of information and low cost made it the best choice for this thesis. The product's focus on providing information for one client at a time was a disadvantage for this research. It was difficult to extract information about several areas in a timely fashion. By limiting the number of areas studied eNeighborhoods was able to provide in-depth and accurate economic demographic data for the areas that comprised the data used in this study.

eNeighborhoods provides several categories of data that are useful to homebuyers and some of which are useful for this study. The data provided can be categorized into three different areas. The first is realty data. Detailed information about almost every home sold in America for the last year is included in eNeighborhoods. This data can be

used to find pricing trends and housing characteristics for several resolution levels. The second area is community features. Data about the area climate, crime rate, and school systems are included for every city in America. The third area is economic and population demographics at the state, county, city, and census tract levels. This is data about the income, age, marital status, education, and employment status of the people that live in a given area. Figure 16 is an example of output from eNeighborhoods from the city of Houston. Collecting data about cities from eNeighborhoods is straightforward.

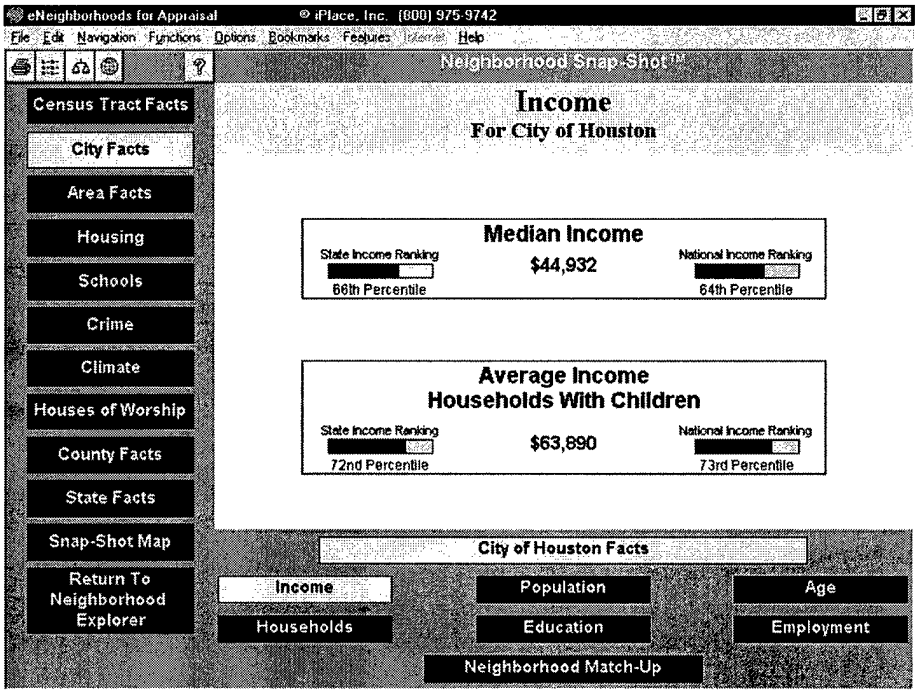


Figure 16 eNeighborhoods City Economic Data

The location finder displayed in Figure 17 allowed the cities included in the study to be located quickly. Once the city was located, the data was manually transcribed into a spreadsheet. The seventeen variables that were collected for each city can be seen in Table 3.

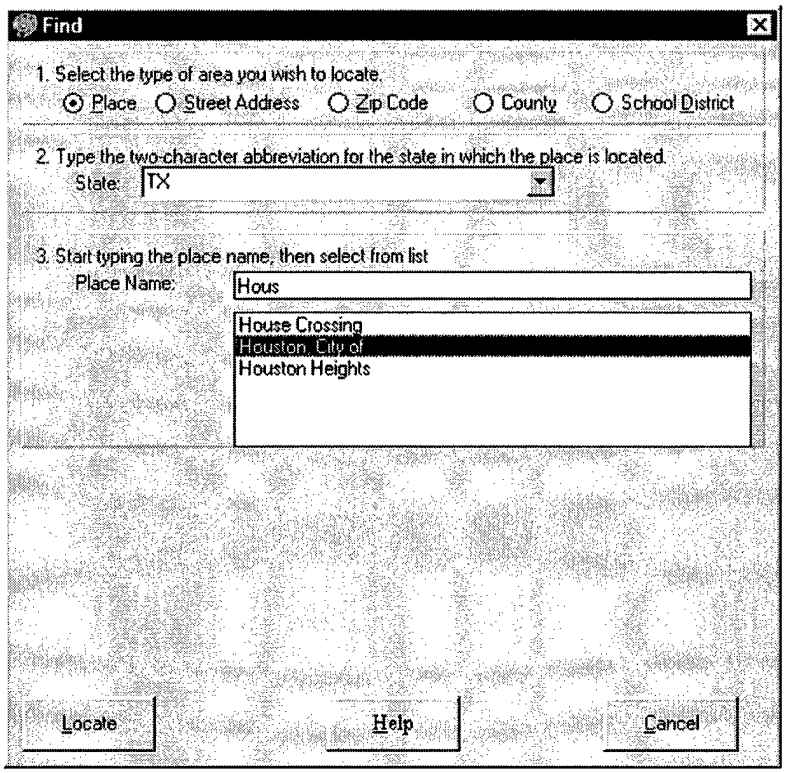


Figure 17 eNeighborhoods Locator

Table 3 Variables Incorporated in This Study

Median Income	Measured per household and not per person or wage earner
Average income for Households with children	
Population density	Annualized over 5 years
Population growth rate	
Average age	The Air Force has age requirements for enlistment
% Population 18-29	
% Households w/children	Martial and parental status may effect values
% Households married	
% High school graduates	Education level of the community
% College graduates	
% Administrative support	Secretarial or clerical workers
% Sales	Retail or wholesale sales
% Management	Supervisor positions
% Professional	Academically skilled labor
% Service	Skilled and unskilled manual labor
% Labor	Manufacturing workers
% Technical	Engineering and electronics manufacturing.

The median income and average income for households with children are data that are gathered by the Internal Revenue Service and reported by eNeighborhoods. Both variables are included in this research because they are a measure of the economic condition of the people that live in the city.

The following variables concern who and how many people live in a city and are measured and maintained by the Census Bureau. The population density and the population growth rate are measurements of the population distribution. Average age and percent population 18-29 are useful to recruiting because there is an age requirement for recruits. The percentage of households married and households with children are indicators of the number of parents with children that live in a city. The percentage of high school and college graduates are measurements concerning the education level of the community.

The last seven variables all deal with the employment of the people in the city. These seven categories are used by the Census Bureau to categorize the different types of labor. Table 3 gave a brief description of each category.

Each of the seventeen variables was collected for all 100 cities. Each group that was studied included 30 cities. Therefore each group matrix had more rows than columns that allows for a statistically valid covariance matrix calculation.

Collecting data for the neighborhood study was more difficult. The same variables that were available for each city, except population density and population growth, were available for census tracts within each city. The Recruiting Service database had zip codes for each recruit but no information that would allow for a direct

link to census tracts. For the large cities that were used in this study census tracts are smaller than zip codes. Each of the values for the 120 zip codes in the study were approximated by averaging the values for the one to four census tracts that were located inside it. An example of how the approximation worked can be seen in Figure 18.

There were two problems that occurred with this approximation. First, the approximation may not include all of the data in the area of the zip code or may include data from another zip code. Second, the census tracts were all equally weighted when in fact they were not equal divisions of the zip code. There was no way to tell what percentage of the zip codes population lived in a census tract with eNeighborhoods.

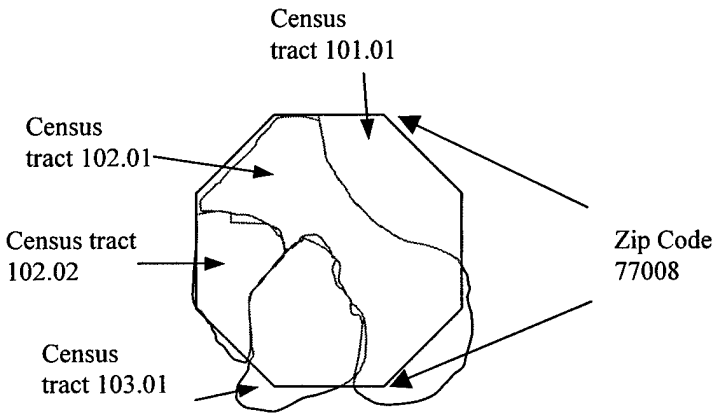


Figure 18 Approximating Zip Codes with Census Tracts

3.5 Calculating Quadratic Discriminant Score Loadings

In Chapter II it was shown that Equation 12 could be used to calculate loadings for groups with equal covariance matrices. Chapter II also suggested that loadings could be calculated for problems that use quadratic discriminant scores to classify the groups.

There are different ways to calculate these loadings. This section demonstrates how the quadratic discriminant (D_q) score loadings were calculated in this thesis.

When using D_q scores to classify the data there are two scores calculated for every object. The first, D_{q1}, is an estimate to the likelihood that the object is from the first group. The second, D_{q2}, is also an estimate, this time for the likelihood that the object is from the second group. The object is assigned to the group with the largest likelihood as seen in Figure 19. In Chapter II it was suggested that the loadings could be calculated one variable at a time with the univariate correlation formula. At this point there are two discriminant scores, D_{q1} and D_{q2}. These scores have to be combined in order for the calculation to be performed. Equation 14 is the composite discriminant score, D_{qc}, formed by subtracting D_{q2} from D_{q1}. Calculating the univariate correlation of each variable with the composite quadratic discriminant score formed the quadratic discriminant loadings.

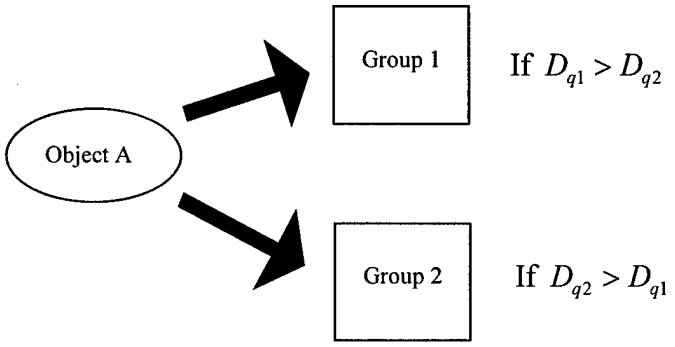


Figure 19 Classifying Data With D_q Scores

$$D_{q1} - D_{q2} = D_{qc} \tag{14}$$

3.6 Variable Reduction

The loadings will show a variable's contribution to the classification. Some of the data provide by eNeighborhoods is useful for classification and some of it is not. If a variable's correlation with the classification scores is low it is probably not important. The true test of its importance is to remove that variable from the classification function and see how the classification accuracy is affected. A goal of this research is to create an accurate and parsimonious classification model. This section explains the procedure, seen in Figure 20 that was used to reduce the number of variables in the classification model.

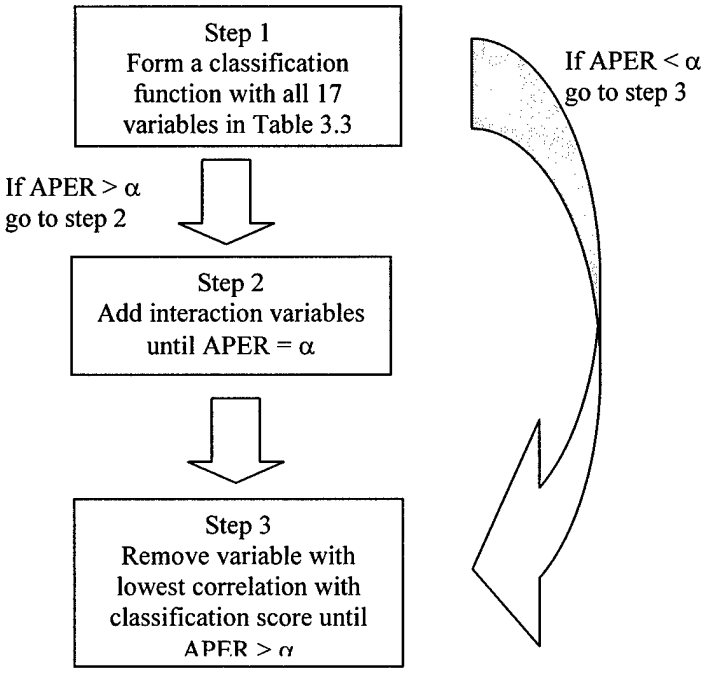


Figure 20 Variable Reduction Process

The first step is of the process is to build a model with all of the available data without interacting any of the variables. It is more difficult to explain variable interaction

than single variables. It is desirable to avoid interactions if an accurate classification function can be formed without them. After the data has been classified the apparent error rate, APER, should be measured. If the error rate is acceptable the researcher should move to step 3. For this thesis, an APER of 0 is possible and only that error rate is acceptable. If the APER is too high to accomplish the goal of the research, then more data is needed to classify the data. If it is not possible to collect more data, it is possible to reduce the misclassification rate by adding interaction variables.

In step two, interaction variables are added to the data until the classification accuracy is acceptable. These variables can be added one at a time, a subset of variables can be added at one time, or the entire set of interaction variables can be added at once. In this thesis a subset of the interaction variables is added each time. In both of the studies the sample size is small. A requirement of discriminant analysis is to have more objects than variables. The subset is chosen in the following manner. One of the single variables that is important to the classification is interacted with the remaining variables. For example, if X_1's loading value has the largest magnitude then the interaction variables $X_1 \cdot X_2, X_1 \cdot X_3, \cdots X_1 \cdot X_{16}, X_1 \cdot X_{17}$ are added. If the APER does not improve, remove the interaction variables before adding any new ones. If the APER = 0, no more interaction variables need to be added, if not then continue adding variables.

Once an acceptable APER is obtained, step 3 can begin. Variables that have low correlation with the discriminant function are removed in a stepwise fashion. Removing too many variables at a time is avoided in this thesis. Because the discriminant scores are a function of data, every time a variable is removed from consideration the discriminant scores change. Therefore a variable's correlation will change through out the variable

reduction process. In this thesis a small number of variables were removed each time and the loadings recalculated.

Variable reduction continues until the APER increases. In the case of this thesis when the APER becomes greater than 0 the process stops. The smallest set of variables that produce an APER of 0 is the set that is used to classify future data.

3.7 Confirming Results

After a set of variables has been selected and a classifying function has been formed, the robustness of that function should be tested. In this thesis, measuring the classification accuracy on independent data will test the robustness of the function.

The method of obtaining independent data used in this thesis is to split the data that has already been collected into training data and holdout data. Training data is the subset of the data that is used to form the classification function. The holdout data is then classified using the function. Comparing the true group affiliation of the holdout data to how the function classified it is an assessment of the accuracy. Classification functions are more accurate if they are formed with as much data as possible. For this reason the training data should be a larger subset of the data than the holdout data. In this research the training subset is at least two-thirds of the original data. Figure 21 demonstrates how training and holdout data is generated in a two-group problem.

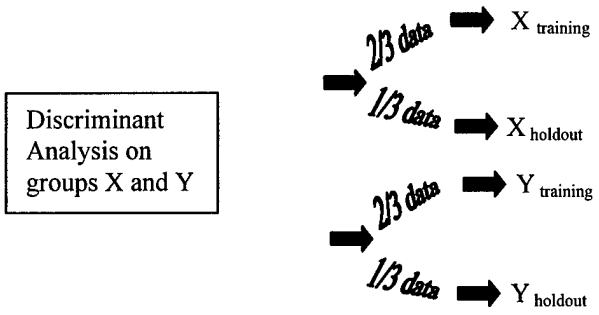


Figure 21 Using Holdout Data for Validation

This chapter discussed how the data from Air Force Recruiting Service was used to define *a priori* groups and how the commercial product eNeighborhoods was used to collect independent demographic data. The last sections of Chapter III specifically laid out how discriminant analysis theory was applied to this research effort. Chapter IV gives the results that resulted from this methodology.

IV. Results

The results in this chapter are the validation of the methodology established by this thesis. The first section displays the results of group formation and the second section explains which classification method was chosen and why. In the third section the results of the strategic study to find indicators of demand across cities. The key variables that act as indicators are identified and their impact on recruiting is explained. In the fourth section, the same results are reported for the tactical study to find indicators of demand within cities. The final section discusses a surprising similarity between Recruiting Service Data and Air Force Reserve Officer Training Corps data.

4.1 Group Formation Results

The group formation methodology explained in Chapter III demonstrated how to define a two-group discriminant analysis study on a continuous variable. In this section the results of applying that methodology on the recruiting data is presented. Two major studies were completed in this thesis: a strategic study identifying demand across cities and a tactical study identifying demand within cities.

The 100 largest cities in the United States were included in the study to identify demand across cities. The cities were ranked according to the number of recruits that came from each city. The cities were divided into three groups as seen in Tables 4, 5 and 6.

Table 4 High Producing Cities

City	Number of Recruits
San Antonio	2161
Houston	1102
El Paso	1090
Chicago	964
Oklahoma City	806
San Diego	729
Colorado Springs	703
Philadelphia	697
Shreveport	697
Miami	695
Las Vegas	634
Tucson	629
Jacksonville	591
Phoenix	571
Virginia Beach	537

City	Number of Recruits
Tampa	507
Cleveland	504
Cincinnati	498
Los Angeles	477
Albuquerque	463
Dayton	449
St Louis	442
Sacramento	436
Memphis	429
Fort Worth	426
Tacoma	426
Detroit	415
Indianapolis	410
Dallas	402
Austin	400

Table 5 Bottom Producing Cities

City	Number of Recruits
Fort Wayne	196
Akron	195
New York	189
Seattle	180
Grand Rapids	179
Lincoln	169
Greensboro	164
Little Rock	161
Garland	156
Jackson	154
Raleigh	148
Nashville	139
Stockton	128
Rochester	124
Anaheim	121

City	Number of Recruits
San Francisco	120
Hialeah	110
Newark	107
Lexington	103
Fremont	98
Des Moines	90
Oakland	74
Boston	71
Madison	70
Jersey City	56
Huntington Beach	56
Santa Ana	54
Glendale	45
Yonkers	35
Alexandria	32

Table 6 Middle Producing Cities

City	Number of Recruits
Buffalo	394
Minneapolis	378
Wichita	378
Omaha	361
Louisville	359
Montgomery	356
New Orleans	352
Honolulu	340
Milwaukee	331
Columbus	329
Spokane	328
Bakersfield	327
Pittsburgh	325
Columbus	305
San Jose	297
Mesa	295
Anchorage	295
Portland	280
Birmingham	269
Corpus Christi	267

City	Number of Recruits
Riverside	263
Kansas City	262
Charlotte	255
Fresno	255
Arlington	255
Mobile	245
Atlanta	242
St Paul	239
Tulsa	238
Norfolk	237
Long Beach	234
Richmond	233
Lubbock	231
Baltimore	228
Toledo	226
St Petersburg	224
Washington	223
Denver	213
Baton Rouge	212
Aurora	207

The cities that produced less than 200 recruits over the last six years made up the "low producers" group. Cities that produced more than 400 recruits made up the "high producers" group. These two groups were used to formulate the classification scores. The remaining cities that produced between 200 and 400 recruits over the last six years were labeled "middle producers" and were not included in the study.

In the study to identify demand within cities, 8 neighborhoods from the largest 15 cities in the United States were used to formulate a classification function. The resulting groups were formed with 60 neighborhoods in each group. The tables containing the

neighborhood data are too large to incorporate in the body of the text. They can be found in Appendix 2.

4.2 Choosing a Classification Method

In Chapter II we presented two criteria for choosing a classification method. Firstly, if the two groups have equal covariance matrices Fisher's method is the correct choice. If the covariance matrices cannot be taken as equal then compute both Fisher's method and the quadratic discriminant method and move to the second criteria. Secondly, compare the error rates of the classifying function for the training data and on independent data. Choose the method with the lowest error rates.

For the data in this thesis, the null hypothesis that the covariance matrices were equal was always rejected. Because of this, Fisher's method was never a clear choice so both classification methods were computed. The apparent error rates from the quadratic discriminant scores were less than, or equal to, the error rates from Fisher's method at every step of the analysis. Therefore, the quadratic discriminant method was used to complete both studies in this thesis.

4.3 Identifying Demand Across Cities

Discriminant analysis techniques identified five variables that provide excellent accuracy on training data and holdout data. The procedure explained in Section 3.7 was successfully executed on this data. In general, variables were added until the apparent error rate, APER, was reduced to zero. After an APER of zero was achieved, variables with low correlation with the composite quadratic discriminant score (Equation 14) were removed in a stepwise manner. The specific steps of the data reduction can be found in

Appendix 3. The indicators of demand that were identified in this study are shown in Table 7.

Table 7 Demand Indicators for Cities

Variable	Loadings	Mean of Bottom Producers	Mean of Top Producers
Median Income Of Households	-.323	45,900	39,200
Average Income of Households with Children (Parent's Income)	-.324	62,270	53,490
% College Graduates	-.408	25.06	20.13
% Workers in Service Jobs	.436	13.9	15.97
Median Income x Parent's Income	-.573	$2.9 \cdot 10^{10}$	$2.1 \cdot 10^{10}$

Four of the five variables, Median Income, Parent's Income, the interaction of the two, and the Percentage of College Graduates, have a negative impact on recruiting as they increase in magnitude. Only the percentage of service workers has a positive impact on recruiting as it increases.

The percentage of college graduates is a measure of the educational level of the community. As the number of people with a college education increases, the inclination to join the military goes down. The percentage of service workers behaves exactly opposite. Both of these variables seem to say that a community with a larger traditional "blue collar" workforce is more favorable to recruiting.

Three of the variables are measures of the income of the population. The variables median income and parent's income suggest that as a community becomes more affluent the less likely people will want to serve in the military. Why is the interaction of the two

variables important? The interaction of the median income of all households with the average income of households with children is the most important variable for classification. The relationship between these two variables is more important than their magnitudes.

To explore this relationship, basic statistics were collected about the variables. The most useful results came from subtracting the two variables to see if the difference between the two variables had a substantial impact. Although the results are not statistically different, it seems that the cities in the bottom producers group have a bigger difference between Parent's Income and the Median Income. Table 8 shows the results of investigating this new variable.

Table 8 Subtracting Median Income from Parent's Income

Group	Average	Standard Deviation
Top Producers		
μ	14,258	5,490
.95 CI	(12,208, 16,308)	(4,373, 7381)
Bottom Producers		
μ	16,383	7,661
.95 CI	(13,521, 19243)	(6102, 10,300)

4.4 Identifying Demand Within Cities

Discriminant analysis techniques were able to identify six variables that provided excellent accuracy on training data and holdout data. The procedure explained in Section 3.7 was successfully executed on this data. In general, variables were added until the apparent error rate, APER, was reduced to zero. After an APER of zero was achieved, variables with low correlation with the composite quadratic discriminant score (Equation

3.1) were removed in a stepwise manner. The specific steps of the data reduction can be found in Appendix 3. The indicators of demand that were identified in this study are shown in Table 9.

The main difference in this study with the previous study to find demand across cities was the performance with holdout data. Variables had to be added back into the model until the classification accuracy on the holdout data was acceptable.

Table 9 Demand Indicators for Neighborhoods

Variable	Loadings	Mean of Bottom Producers	Mean of Top Producers
Parent's Income	.273	85,980	54,475
% High School Graduates (HS)	-.035	70	71.32
% College Graduates (CG)	.365	29.8	16.4
% Administrative Support Workers	-.605	15.4	20.4
High School • College Grads	.369	2,534	1,605
College Grads • Admin Support	-.171	613	668

When just using the training set, the data can be reduced to just two variables. When just using the interaction of % high school graduates and % college graduates and the interaction of % college graduates and % administrative works, the APER remains 0. Unfortunately, the classification rate on the holdout data is only 50%. To improve the performance of the classification function on independent data, variables were added back to the model in the order that they were removed. The final set that provided an APER of 0 and 100% classification accuracy on holdout data can be seen in Table 9.

Three of the variables have a positive effect on recruiting and three variables have a negative effect on recruiting. The percentage of administrative support workers is the most important variable. As the percentage goes up, the attitude towards the Air Force improves. Administrative support workers are secretaries and clerical workers. Another way to think of them is as the people that work for professionals. The interaction between the percentage of college graduates and administrative workers also has a positive effect on recruiting as it increases. For the neighborhoods in the Bottom Producers category the college graduates and administrative workers variables were negatively correlated. In the Top Producers category they were almost uncorrelated. This can be seen in Table 10.

Table 10 Correlation of College Grads with Administrative Support

Group	Correlation
Bottom	-.376
Top	-.089

The last variable that shows a positive contribution is the amount of high school graduates. The Air Force has strict requirements that recruits have a high school degree. The variable by itself does not have that much importance, but its importance can be seen when it is interacted with the percentage of college graduates.

The interaction of college graduates with high school graduates has a negative effect on recruiting as its magnitude increases. More specifically when the number of high school graduates is high and the number of college grads is low recruiting improves. This is shown graphically in Figure 22. The percentage of college graduates by itself has a negative effect on recruiting. In addition, the magnitude parent's income also has a

negative effect. Both of these variables show that the affluence of a neighborhood is an indicator of attitude towards enlisting in the Air Force.

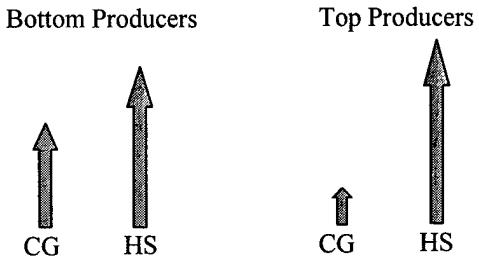


Figure 22 Interaction Between College Grads and High School Grads

4.5 Importance of Hierarchy

In both of the studies conducted in this thesis, the importance of hierarchy was very clear. Hierarchy is the practice of including the lower order terms that make-up the interaction variables in the model (Neter *et al*, 1996). In both of the models created in this thesis, an APER of 0 on the training set could be maintained if the lower order terms were removed from the model. The reduced models consistently performed poorly on holdout data. The classification function became much more robust when the lower order variables were reincorporated in the model.

4.6 Similarity with AFROTC data

The Air Force Reserve Officer Training Corps, (AFROTC), maintains a database similar to the Recruiting Service Database. AFROTC fields a nationwide recruiting force to find people to join the Air Force. There main difference is AFROTC offers college scholarships to high school students. AFROTC has made available data about the

scholarship applicants from the last two years. This was not enough data to determine demand within cities, but there was enough data to complete the study to find demand across cities. An interesting thing happened. When the groups were defined for high demand cities and low demand cities, the groups were almost exactly the same as when they were defined from the Recruiting Service data. Because the groups that were defined contained mostly the same data, the important variables were identical.

This is a surprising result. It would seem that Recruiting Service and AFROTC recruit from two different populations. Recruiting Service traditionally recruits people who are seeking vocational work and AFROTC recruits people seeking a management position requiring a college education. This data shows that they are competing over the same population.

The results of thesis show that the data collected can successfully discriminate good and bad locations for recruiting in large urban areas. In the final chapter these results are summarized and future applications of this research are discussed.

V. Conclusions and Recommendations

5.1 Impact of Indicators

The indicators that were identified in both the strategic and tactical study appears to have uncovered a common theme. People do not want to join the Air Force if they perceive their standard of living will decrease. Military Recruiters should be aware of how these indicators effect potential recruits perceptions and think of strategies to use them to their advantage.

The percentage of the community that is college educated seriously affects Air Force Recruiting. College enrollment and the opportunity for financial aid are steadily increasing. Since the Air Force has many programs that help airmen earn college degrees this is a problem that can be combated. Recruiters and advertising should emphasize the true costs of college and the programs that the Air Force offers to educate airmen.

The occupational make-up of a community has a big impact on recruiting productivity. This research found that communities with a larger percentage of blue-collar workers and administrative workers are more conducive to recruiting. Advertising directly to parents that work in these career fields appears to be one way that these indicators can be used to increase recruiting productivity.

The indicators that were identified are very useful for finding communities that produce individuals that are receptive to joining the Air Force under the current recruiting policies. There is a flip side to the coin. These indicators also identify the areas of the population that the Air Force's message is not reaching. If the Air Force is to continue to meet its personnel needs with an all-volunteer force, it must find ways to expand that part of the population that is receptive to service in the military.

5.2 Possible Applications

Air Force Recruiting Service will face new short and long-term recruiting challenges in the future. This thesis investigated how to find productive recruiting areas amongst and within large cities. A methodology to do this was developed and tested. There are other communities which can be studied.

The Air Force recruits people from rural, suburban, and urban areas. This methodology can be used to understand the difference between recruits that came from these different areas. These results could identify different population characteristics that recruiters could learn to incorporate in business practices. The results of such a study could lead to specialized recruiting tools that meet the special needs of recruiters from the three different areas.

Sometimes it is not enough for the Air Force to recruit a large number of people, at times the Air Force needs people with specific skills that can fill critically unmanned career fields. One current example is the Air Force's under manning in aircraft maintainers. The Air Force has increased the bonus for new recruits that choose to be maintainers. The problem is that a recruit must meet certain minimum standards to be qualified to be a maintainer. This methodology can potentially identify areas where enlistees are more likely to be qualified.

5.3 Suggestions for Future Research

This research was successful, but there are some areas that could be improved or were not explored because of time constraints. There are two areas that need attention,

finding a faster way to collect demographic data, and measuring a greater variety of variable types.

eNeighborhoods is an accurate and flexible platform for collecting some demographic data. Unfortunately, using eNeighborhoods is not a timely enough method for gathering demographic data across a large number of areas. Future research would benefit by identifying a new source for collect data.

When searching for a data source, the researcher should choose a source that incorporates a large selection of variables. This research identified that the interaction between the overall income and the income of people with children. Variables such as percentage of single-earner married households or the percentage of single mothers might bring further understanding to future research.

In conclusion, this thesis successfully identified variables that can be used to find productive recruiting areas in metropolitan areas and avoid nonproductive areas. The results not only details the characteristics of the part of the population the Air Force has done well with, but also details the characteristics of the people the Air Force will have to improve its image with.

Appendix I: List of Cities

Top Producers City	Count Of Recruits
San Antonio	2161
Houston	1102
El Paso	1090
Chicago	964
Oklahoma City	806
San Diego	729
Colorado Springs	703
Philadelphia	697
Shreveport	697
Miami	695
Las Vegas	634
Tucson	629
Jacksonville	591
Phoenix	571
Virginia Beach	537
Tampa	507
Cleveland	504
Cincinnati	498
Los Angeles	477
Albuquerque	463
Dayton	449
St Louis	442
Sacramento	436
Memphis	429
Fort Worth	426
Tacoma	426
Detroit	415
Indianapolis	410
Dallas	402
Austin	400

Middle Producers City	Count Of Recruits
Buffalo	394
Minneapolis	378
Wichita	378
Omaha	361
Louisville	359
Montgomery	356
New Orleans	352
Honolulu	340
Milwaukee	331
Columbus	329
Spokane	328
Bakersfield	327
Pittsburgh	325
Columbus	305
San Jose	297
Mesa	295
Anchorage	295
Portland	280
Birmingham	269
Corpus Christi	267
Riverside	263
Kansas City	262
Charlotte	255
Fresno	255
Arlington	255
Mobile	245
Atlanta	242
St Paul	239
Tulsa	238
Norfolk	237
Long Beach	234
Richmond	233
Lubbock	231
Baltimore	228
Toledo	226
St Petersburg	224
Washington	223
Denver	213
Baton Rouge	212
Aurora	207

Bottom Producers City	Count Of Recruits
Fort Wayne	196
Akron	195
New York	189
Seattle	180
Grand Rapids	179
Lincoln	169
Greensboro	164
Little Rock	161
Garland	156
Jackson	154
Raleigh	148
Nashville	139
Stockton	128
Rochester	124
Anaheim	121
San Francisco	120
Hialeah	110
Newark	107
Lexington	103
Fremont	98
Des Moines	90
Oakland	74
Boston	71
Madison	70
Jersey City	56
Huntington Beach	56
Santa Ana	54
Glendale	45
Yonkers	35
Alexandria	32

Appendix II: List of Neighborhoods

1. New York

Zip Code	Recruits
10027	14
10034	10
10032	13
10029	17
10004	2
10011	1
10012	1
10044	1

2. Los Angeles

Zip Code	Recruits
90016	21
90043	18
90011	18
90045	22
90036	3
90012	3
90049	1
90056	2

3. Chicago

Zip Code	Recruits
60620	52
60636	27
60651	27
60647	35
60601	3
60604	2
60605	2
60664	1

4. Houston

Zip Code	Recruits
77089	39
77060	22
77080	20
77015	44
77006	2
77218	2
77030	2
77056	1

5. Philadelphia

Zip Code	Recruits
19143	40
19124	31
19132	29
19120	39
19103	1
19137	5
19127	2
19112	1

6. San Diego

Zip Code	Recruits
92115	26
92154	75
92104	28
92126	92
92108	5
92172	1
92136	1
92106	3

7. Detroit

Zip Code	Recruits
48227	32
48224	51
48205	31
48228	47
48233	1
48217	5
48208	5
48216	1

8. Dallas

Zip Code	Recruits
75224	19
75243	23
75228	23
75240	18
75203	2
75233	3
75207	1
75210	2

9. Phoenix

Zip Code	Recruits
85037	47
85023	34
85022	22
85032	41
85086	1
85001	1
85045	2
85034	4

10. San Antonio

Zip Code	Recruits
78240	80
78223	67
78244	74
78228	74
78289	1
78203	4
78215	2
78265	1

11. San Jose

Zip Code	Recruits
95136	19
95148	24
95111	21
95127	22
95116	4
95110	3
95134	3
95117	4

12. Baltimore

Zip Code	Recruits
21215	24
21229	20
21206	27
21224	19
21201	1
21250	1
21231	3
21209	2

13. Indianapolis

Zip Code	Recruits
46260	16
46241	19
46227	35
46201	16
46231	4
46216	2
46225	2
46266	1

14. San Francisco

Zip Code	Recruits
94134	9
94102	8
94112	24
94110	10
94115	2
94133	2
94123	2
94139	1

15. Jacksonville

Zip Code	Recruits
32210	67
32211	27
32208	32
32209	29
32204	3
32297	1
32214	1
32212	1

Appendix III: Variable Reduction

Step	Data Changed	APER	Low Correlation	Notes
1	All 17 original variables included	Fisher = .15 D_q = .033	Households w/ children 0 % High school graduate -.089 % Administrative -.053 % Labor .054	The first attempt was to remove "unimportant" variables and see if a good classification could be maintained
2	The Four low correlation variables removed	Fisher = .17 D_q = .067		Classification error rate doubled but is still acceptable. Try new approach
3	Add interaction variables: Median Income (MI) with all others	Fisher = 0 D_q = 0	Average Age .056 MI* Average Age -.026 MI*% Sales .015 MI*%Service -.058 MI*%Technical 0	MI*Population density was removed because of dependency problems
4	The Five low correlation variables removed	Fisher = 0 D_q = 0	% Management -.087 MI* Management -.108	Remove variable and interaction to maintain hierarchy
5	The two low correlation variables removed	Fisher = 0 D_q = 0	MI*%College Grad -.121	
6	Low correlation variable removed	Fisher = 0 D_q = 0	% Professional -.131	
7	Low correlation variable removed	Fisher = 0 D_q = 0	% 18-29 -.133	
8	Low correlation variable removed	Fisher = 0 D_q = 0	%Households Married .134	
9	Low correlation variable removed	Fisher = 0 D_q = 0	% Sales 0	
10	Low correlation variable removed	Fisher = 0 D_q = 0	Population growth rate .054	
11	Low correlation variable removed	Fisher = 0 D_q = 0	% Technical .179	
12	Low correlation variable removed	Fisher = 0 D_q = 0	Population density -.265	
13	Low correlation variable removed	Fisher = 0 D_q = 0	Median Income -.343	Violates hierarchy but error rate = 0
14	Low correlation variable removed	Fisher = 0 D_q = 0	Parent's Average Income -.387	Violates hierarchy
15	Low correlation variable removed	Fisher = .22 D_q = .183		Include all data from step 14.

Bibliography

1. Andrews, D. F., R. Gnanadesikan, and J. L. Warner. "Methods for Assessing Multivariate Normality", Multivariate Analysis-III "Proceedings of the Third International Symposium on Multivariate Analysis held at Wright State University, Dayton, OH, 19-24 June 1972". Academic Press. New York, 1973.

2. Bauer, K. W. "OPER685, Applied Multivariate Data Analysis," Fall 2000. Air Force Institute of Technology, OH.

3. Cheezum, Debbi. "Youth Marketing", MarketSource Corporation Teen Marketing, 30 March 2000, http://www.marketsource.com/teen/main_teen.asp.

4. Chernoff, Herman. " Some Measures for Discriminating between Normal Multivariate Distributions with Unequal Covariance Matrices", Multivariate Analysis-III "Proceedings of the Third International Symposium on Multivariate Analysis held at Wright State University, Dayton, OH, 19-24 June 1972". Academic Press. New York, 1973.

5. Cordeiro, James D., Jr., and Mark A. Friend. Using Simulation to Model Time Utilization of Army Recruiters. Master's Thesis, AFIT/GOR/ENS/98M-06, AFIT/GOR/ENS/98m-12. Air Force Institute of Technology, Wright-Patterson AFB OH, March 1998.

6. DeHaan, Laura. "NDSU Research: Community Influence Important in Adolescent Development", News for North Dakotans, 4 February 1999, http://www.ext.nodak.edu/extnewsrelease/1999/020499/02ndsure.htm.

7. Dillon, William R. Multivariate Analysis, Methods and Applications. John Wiley & Sons. New York, 1984.

8. "eNeighborhoods for Realty", Iplace Professional Services, 1 October 2000, http://www.iplacepro.com/realty.asp

9. Giri, Narayan C. Multivariate Statistical Analysis. Marcel Dekker, Inc. New York, 1996.

10. Hafemeister, Rod. "Recruiting to face 'continuing challenges'". *Air Force Times*, 14 August 2000.

11. Klopper, Susan M. "Unearthing Market Research: Get Ready for a Bumpy Ride," *Searcher*, 8 (3), March 2000, pag. http://www.infotoday.com/searcher/mar00/klopper.htm – Article discussing how market research is gathered.

12. Longhorn, David C. Using Simulation to Model an Army recruiting Station with Seasonality Effects. MS Thesis, AFIT/GOR/ENS/00M-18. Graduate School of Engineering of the Air Force Institute of Technology, Wright-Patterson AFB OH, March 2000.

13. Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. Applied Linear Statistical Models, 4th edition. Irwin. Chicago, 1996.

14. O'Connor, Larry. "It's a tale of two districts: Livonia, Clarenceville," *Observer-Eccentric Newspapers*, 28 March 2000, http://www.observer-eccentric.com/Livonia50/services/publicschools.htm.

15. O'Fearna, Frank C, Captain, USAF. Chief, Market Research Systems and Training, Air Force Recruiting Service, Randolph AFB TX. Telephone interview. 10 September 2001.

16. Palmer, Jennifer. "I'd Rather Go Short Than Go Stupid". *Air Force Times,* 8 May 2000.

17. Palmer, Jennifer. "New Logo Will be Phased in Over Years". *Air Force Times,* 31 June 2000.

18. Rolfsen, Bruce. "Forced Duty, Air Force to Draft NCOs into Ranks of Recruiters". *Air Force Times,* 11 December 2000.

19. Rolfsen, Bruce. "5 Level Crisis". *Air Force Times,* 18 December 2000.

20. Snee, Ronald D. "In Pursuit of Total Quality", Quality Progress, April 1986.

21. Vanfossen. "The Big Picture of Economic Restructuring", Institute for Teaching and Research On Women (ITROW) News, Fall/Winter 1996, http://www.towson.edu/~vantoss/review.htm.

22. Wackerly, Dennis D., William Mendenhall III, and Richard L. Scheaffer. Mathematical Statistics with Applications, Duxbury Press, New York, 1996.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
20-03-2001	**Master's Thesis**	July 2000 – Mar 2001

4. TITLE AND SUBTITLE
IDENTIFING DEMAND INDICATORS FOR AIR FORCE RECRUITING SERVICE WITH DISCRIMINANT ANALYSIS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Williams, Jason, L., 1st Lieutenant

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GOR/ENS/01M-18

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Frank O'Fearna, Captain, USAF
Chief, Market Research Systems and Training
AFRS/RSOAM
550 D Street West Ste. 1
Randolph AFB, TX 78150-4527 DSN:487-2331

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A changing public disposition towards military service has all four military branches rethinking recruiting practices. This Air Force is reacting to this new recruiting climate by increasing the bonuses for new recruits, pumping up advertising budgets, and bolstering recruiting personnel levels.

This thesis provides a new tool for assessing how to allocate these new resources. Discriminant Analysis is used to identify population characteristics that categorize recruiting locations. A methodology is constructed that can discriminate between communities where interest is high in military service and where recruiting efforts will not be productive.

15. SUBJECT TERMS
Discriminant Analysis, Feature Selection, Recruiting, Hierarchy

16. SECURITY CLASSIFICATION OF: UNCLASSIFIED			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kenneth W. Bauer
U	U	U	UU	71	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-6565, ext 4328 (Kenneth.Bauer@afit.edu)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18